# Research on Big Data-Driven High-Risk Students Prediction

Yu Xiaogao
School of Information Management and Statistics
Hubei University of Economics
Wuhan, China
e-mail: tecom_sam@163.com

Peng Ruiqing
School of Information Management and Statistics
Hubei University of Economics
Wuhan, China
e-mail: 469554298@qq.com

***Abstract*—Big data technology is used to integrate, clean up and analyze the data of student management system, educational administration system, and campus card consumption system and so on in this paper. The characteristics of high risk students are extracted and selected, and the prediction model is constructed, which can be used to predict the high risk students scientifically, reasonably and effectively, we can take care of these students and take effective measures to help them improve. The massive data accumulated in Colleges and universities can be effectively utilized by this study, managers have a more detailed understanding of colleges and universities and students, the teaching effect is improved, the risk of student drop-out is reduced. Therefore, this study has strong theoretical value and important practical value.**

***Keywords-big data; prediction;student; model***

## I. INTRODUCTION

With the development of smart phones, the Internet of things and cloud computing, education field accumulated huge amounts of data, education big data analytics is imminent. However, the current research on education big data is not enough, as the data is small, the processing method is simple and the scope of application is narrow. Big data technology can discover hidden patterns and knowledge from massive data [1], it has been applied in biological, financial and e-commerce and other fields. In recent years, Educational Data Mining in big data (EDM) has become a new research field driven by the application of educational information, distance education and Web2.0 [2].EDM is an interdisciplinary subject that spans many fields such as pedagogy, psychology, computer science and Mathematics [3].With the help of educational expertise and big data technology, EDM can discover new knowledge in education and help educational institutions to achieve their educational goals more efficiently. The application of EDM is very extensive, such as student performance prediction, student modeling and adaptive learning, in which student performance prediction is one of the earliest and most important research directions of EDM [4-6], it can predict or judge the learning effect of the students by using the data generated by them in the learning activities. The data sources of high risk students are very rich, and it can be the basic personal information, cultural background, social background, family economic condition, psychological condition, educational level, previous learning situation and even interpersonal relationship [7-11].However, most of the current EDM research results are based on a single data source, data structure and data processing methods are relatively simple.

In this paper, high-risk students are more likely to have multiple classes failing in the final exam, leading to repeat or even drop-out students. Earlier in each semester, the discovery of high-risk students can remind counselors and instructors to intervene and help students in a timely manner, reducing the risk of student dropout. In this paper, we identify high-risk students using existing data from universities. In particular, this paper studies how complex data from multiple information systems can be used to predict whether a student will fail multiple courses at the end of the semester. The key to solving this problem is to extract accurate and appropriate student characteristics from multiple complex data sources.

## II. REVIEW

High-risk students prediction can help students to find problems and improve learning methods or strategies, but also help teachers and counselors go over the overall situation of students, they can help them timely. In order to strengthen study style construction, improve students' learning initiative and ensure the quality of personnel training, it is necessary to predict high risk students. High-risk students prediction has always been the focus of educational science research, according to the different teaching environment, the research status of high risk students is introduced as follows.

### *A. Closed Teaching*

The closed teaching system is mainly a stand-alone learning system and a management information system based on C/S (Client/service) structure. The amount of data is small for such systems. Such as Natek found the key factors that affect the rate of students' curriculum passing through the decision tree algorithm to analyze the data of University Information System and succeeded in predicting the final exam scores[12].

### *B. Open Teaching*

Open teaching environment allows students to communicate and collaborate with each other, the most typical system is the intelligent tutoring system (ITS).Lara and others established the reference model by using the historical data of ITS courses, using the model to identify

978-1-5090-4499-3/17/$31.00 ©2017 IEEE

whether a student can successfully complete the course[13].Romero and others analyzed the students use the Learning Forum and successfully predicted the final score of students by using classification and clustering algorithm[7].

### C. New Teaching Environment

In recent years, many new teaching environments has emerged with the rapid development of technical means, for example, the teaching environment based on the game, social networking, smart mobile devices and augmented reality technology and the massive open online course. These new teaching environments have also led to some interesting topics, such as the [12] to predict the students' learning by 47 college students majoring in computer in computer game data, the accuracy rate is more than 85%.

Compared with foreign research, domestic research started late in the field, and there is a big gap in the breadth and depth of the study [11,14]. In the past ten years, some progress has been made in the research of education and teaching data in China, but there are still some deficiencies [2-6]. Mainly reflected in three aspects: first, innovation is not strong, many results of the study are reviews of foreign studies, tracking and improvement; second, the depth of technology is not enough, many of the research results published in educational journals; third, the research scope is narrow, the research results mainly concentrated in the intelligent tutoring system and personalized learning [2, 8], and there are few studies on the prediction of high risk students in china.

At present, the existing problems are mainly manifested in the following three aspects: first, the data source is relatively simple. At present, the research is aimed at a simple data set, the data source is relatively single, the data is only from a system or a course [9-11], at the same time, the amount of data is small, the amount of data is generally not more than 1MB [5-8]. Second, the data is relatively easy to come out. From the published research results, the data involved in the structure and content is relatively simple, with less noise data, complex data cleaning is not required [9]. Third, the scope of application is narrow. The existing research work is to model according to a specific course data, which actually limits the scope of the research results [4, 7].

In this paper, big data technology is adopted and the massive data of multiple data sources are utilized, the model features are extracted by well-designed data processing methods, and a more general high-risk student prediction model is constructed, which can be used to predict the high risk students scientifically, reasonably and effectively, we can take care of these students and take effective measures to help them improve.

## III. METHODS

It is proposed to use iterative development to study big data-driven high-risk students prediction, the final system starts with a simple prototype system, and then evolves through several intermediate systems. Each intermediate system will introduce new features on the basis of the previous system, including data acquisition, feature extraction, model building and system development.

The study of high risk students prediction based on big data mainly includes the extraction of students' characteristics, the construction of predictive models and the development of predictive system.

Figure 1 shows the total structure of research, At the same time, the hierarchical relations of the three parts are pointed out: Feature extraction is the integration of the various data sources and analysis, to extract a number of student characteristics, it is the foundation of the whole study; Model building is based on the characteristics of students; System development is to achieve the prediction model. Each section is described as follows.

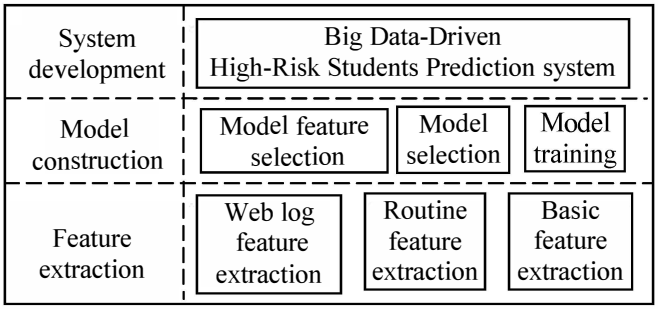


Figure 1. Total structure.

### A. Students Feature Extraction

The feature extraction is to collect and process student-related data, to provide the required student characteristics for the prediction model. According to the different data sources, it can be divided into basic feature extraction, routine feature extraction and network log feature extraction.

First, the basic feature extraction. Basic characteristics include the basic information of students (such as professional, grade, gender, ethnic, regional, etc.) and performance information (such as college entrance examination scores, subjects, failing subjects). This information mainly comes from the student management information system and educational administration management system. This section includes the basic information collection, processing, conversion and statistics, mainly using conventional processing methods. However, some missing and contradictory data need to be dealt with specifically.

Second, routine feature extraction. The routine feature is used to calculate the regularity of students' daily habits. The data source is mainly the students' campus card information, including three meals consumption record, using the hot water record and the access control system records. The corresponding algorithm is designed to extract their routine patterns and calculate the degree of routine regularity.

Third, web log feature extraction. Web log features are used to determine the time taken by students to access various types of web sites.

Step 1. Tens of thousands of Web site domain name is divided into hundreds of small categories, dozens of major categories and several major categories. Site classification needs to be combined with known classification table, artificial classification and computer-aided classification.

Step 2, special site browsing time estimates. Some special types of Web sites (such as video sites, gaming sites and learning websites) have a great impact on the performance of high risk students prediction, so the estimation accuracy of their browsing time is very high. This kind of website can be divided into the state class website and stateless class website.

Step 3, other web browsing time estimates. Other sites other than special sites have little influence on the prediction performance, so a rough time estimation algorithm can be used.

### *B. Prediction Model Construction*

Prediction model construction is to select the appropriate students' characteristics and classifier model to train and challenge the classifier through historical data, and finally builds a high-risk students prediction model with good performance. It is divided into model feature selection, model selection and model training.

First, model feature selection. a few of the features are used in the prediction model in the feature extraction stage, which is to extract the students characteristics in the imbalanced samples.

Second, model selection. Model selection is to select the appropriate classifier to predict high-risk students. At present, there are more than 20 kinds of commonly used classifier algorithms, such as neural network, Naive Bayes, SVM, Logistic regression and so on. Even the same classifier may have different parameter settings. This part will select a few classifiers and parameter settings as a candidate model for the next model training.

Third, model training. In this part, several candidate models are trained by historical data. Finally, the best prediction model is selected, which needs to solve the problems of training mode, massive data processing and sample selection.

### *C. Prediction System Development*

Finally, the prediction model is converted into a software system to test and improve the theoretical model, and it provides services for students and teachers. This part has completed the technology selection, the software architecture design, the system function design, the multi data source data interface design and the software process management and so on. The main software functions include data acquisition, model implementation, user management and e-mail notification.

## IV. ARCHITECTURE

The architecture of big data-driven high-risk student prediction is shown in Figure 2, it is divided into four layers, they are data acquisition and presentation layer, data application layer, data storage layer, cloud computing platform layer. The data application layer is divided into four parts, they are data sharing management, data exchange center, platform monitoring and high-risk students prediction analysis. Data warehouse technology is used to construct the data center; data mining in big data and other technologies are used to achieve big data-driven high-risk student prediction function. Data service function can provide a variety of adapters and service bus. Student management system, educational management system, campus card system and other subsystems can provide massive data the data exchange center stores the data in the data warehouse located in the shared data center according to their different resource forms in data exchange technology.

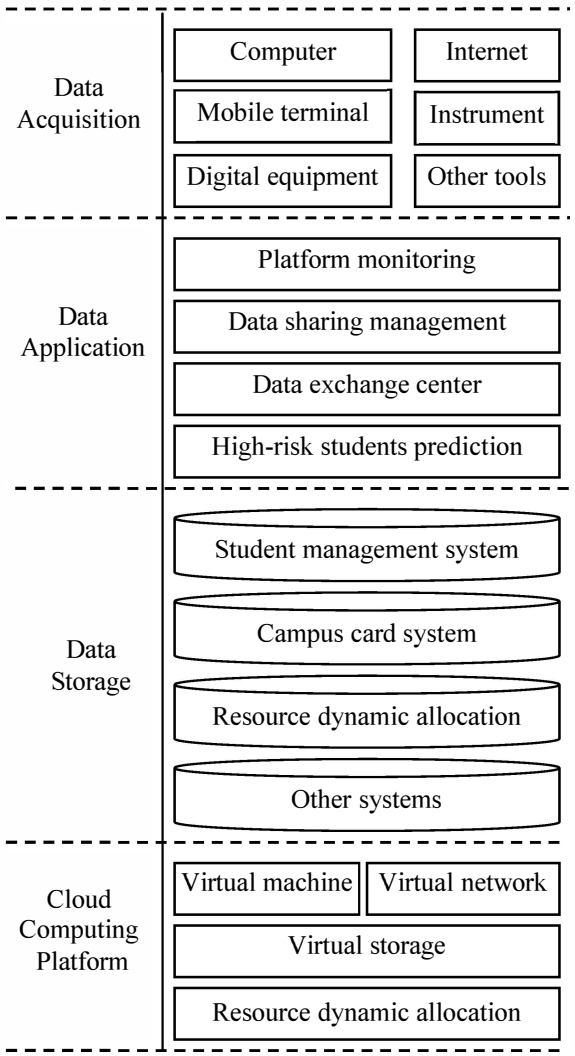


Figure 2. Architecture.

In the data acquisition layer of the high-risk students prediction system, the students' data are collected by computer, mobile terminal, digital equipment, instrument and other tools; the basic platform of the system based on the distributed processing of cloud computing is constructed, which provides the most basic physical platform to support the application of big data technology; In the data storage layer and data application layer, various types of unstructured and structured data is stored in HDFS technology of HADOOP, these data are mined and analyzed using big data mining techniques to provide high-risk students prediction services for teachers, counselors, university administrators and parents.

## V. KEY TECHNOLOGY

This paper focuses on data cleaning and feature selection. The existing system includes a large number of noise and missing data, and the behavior of students is complex, It is necessary to analyze the original data with the combination of computer science, statistics and pedagogy, involving data collection, integration, transformation, research, experiment, statistics and visualization and other research methods. Two key problems need to be resolved in this paper: The first question is to estimate the time that a student visits a stateful Web site based on the raw web log file. The second problem is to select a few representative characteristics from the hundreds of attributes of the imbalanced sample.

### *A. Time Estimation of Stateful Web Sites*

A stateful Web site means that a series of HTTP requests sent by user to the site are controlled by a hidden state mechanism, with dependencies between different types of requests. Typical stateful websites include online learning sites and document sites. For example, the most common types of requests are HTMLView, Jasonload, Browse, GetWords, PageViewTime and other types in Baidu library website. By observing the order of different types, Baidu library request state machine can be derived. The time of user browsing the web site can be estimated more accurately by combining the state machine and the access sequence.

Step 1, the problem will be regarded as the inverse problem of generative grammar, generative grammar is known grammar rules (state machine), these rules are used to judge whether a sequence of symbols (state sequence) is legal; and the problem is a number of known state sequences. Thus, a state machine is constructed.

Step 2, first of all, the statistics on the state machine, remove those few frequency state, because the presence of such states will greatly increase the complexity of the model, and the removal of such states has little effect on the estimation of time. Then, the hidden Markov chain is constructed according to the state sequence.

At present, the state machine model of Baidu library has been successfully established in step 2.

### *B. Feature Selection of Imbalanced Samples*

The proportion of high-risk students (positive samples) to all students is very small, usually below 10% and some even less than 1%.Moreover, the absolute number of high-risk students is usually very small, and the number of high-risk students is below 10 in most majors. If the traditional machine learning algorithm is used to construct the prediction model, because the number of low risk students is much higher than that of high risk students, the traditional model will tend to predict the low risk students. At the same time, a large number of high-risk students will be masked by the characteristics of low-risk students; the predicted effect can not be achieved.

Analysis of big data shows that the time of high risk and low risk students visiting each site is subject to exponential distribution. Therefore, a key scientific problem to be solved in this paper is how to select the most relevant characteristics of the high risk students from the imbalanced samples for exponential distribution characteristic.

In this paper, it is shown that the better feature subset can be obtained by adding the constructed positive samples. The risk type of students is Y, which is 1 (representing a high risk student) or 0 (representing a low-risk student), and the proportion of high-risk students was $P_h$, and the proportion of low-risk students was $P_i$, then $P_k(Y) = p_k$, $D(Y) = p_k p_i$ .Set X for student characteristics, are subject to exponential distribution:

$$\int(x \mid \mu) = \begin{cases} \frac{1}{\mu}\theta^{\frac{k}{x}} & x > 0 \\ 0 & x \le 0 \end{cases}$$

According to the exponential distribution, the mathematical expectation and variance of X are $E(X) - \mu$ , $D(X) - \mu^2$ .

It is assumed that the mathematical expectation of high risk and low risk students on characteristic X is $\mu_h$ and $\mu_i$, then it can be deduced that the Pearson correlation coefficient *CC* of the characteristic X and the risk type Y is equal to

$$CC = \frac{\sqrt{p_h p_i}\,(\mu_h - \mu_i)}{\sqrt{2\mu_h^2 p_h + 2\mu_i^2 p_i - (\mu_h p_h + \mu_i p_i)^2}}$$

$$\text{or} \quad CC = \frac{\sqrt{p_h p_i}\,(r-1)}{\sqrt{2p_h r^2 + 2p_i - (p_h r + p_i)^2}}$$

$r = \mu_h / \mu_i$ is the ratio of the mathematical expectation of positive and negative samples.

From the calculation, the absolute value of the correlation coefficient increases with the positive sample ratio for the negative correlation characteristics (r<1); For positive correlation characteristics (r>1), the correlation coefficient decreases slowly after a significant increase, but the correlation coefficient after adding the sample is usually greater than the correlation coefficient before the sample is added. This shows that the characteristics of positive samples masked by negative samples can be presented by adding a positive sample constructed.

## VI. APPLYING EFFECTS

The system has been tested in Hubei University of Economics. At present, we have collected the data of student management system[15], educational administration system[16], campus card system and student attendance system. Compared with the actual situation of the students, we found that the bigger the data volume and the more data sources, the better the results. As shown in Table 1 and Figure 3.

TABLE I. THE EFFECT OF OF DATA VOLUME AND DATA SOURCE IN HIGH-RISK STUDENT PREDICTION SYSTEM

| Data volume \ Data Sources | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 500 | 0.652 | 0.768 | 0.839 | 0.927 |
| 1000 | 0.743 | 0.782 | 0.857 | 0.935 |
| 1500 | 0.761 | 0.792 | 0.864 | 0.946 |
| 2000 | 0.789 | 0.823 | 0.875 | 0.958 |
| 2500 | 0.813 | 0.847 | 0.882 | 0.966 |
| 3000 | 0.832 | 0.871 | 0.891 | 0.979 |
| 3500 | 0.847 | 0.889 | 0.902 | 0.982 |
| 4000 | 0.854 | 0.897 | 0.913 | 0.991 |

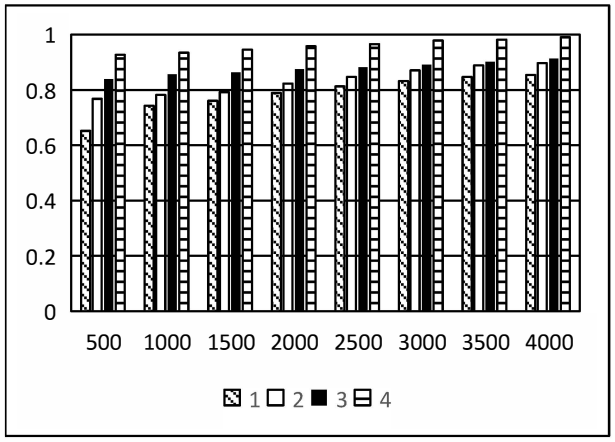


Figure 3. The effect of of data volume and data source in high-risk student prediction system.

## VII. CONCLUSION

Big Data-Driven High-Risk Students Prediction can help students identify problems as early as possible, improve their learning methods or strategies, it can also help teachers or counselors understand the students' overall situation and provide timely help to high-risk students.

Big data technology applied to the education of large data processing to predict high-risk students according to the student's usual discipline, academic performance, thinking, etc... The research of this paper can broaden the research methods and application scope of education big data, so it has strong theoretical value and important practical value.

Big Data-Driven High-Risk Students Prediction is a new subject, and there are still some problems to be solved. The research of this paper is not only beneficial to the teaching work of the school, but also can promote the development of the EDM discipline; it plays a positive role in scientific research and social development. It provides a foundation for further research.

## ACKNOWLEDGMENT

This paper is supported by 2016 Hubei Province Education Science Planning Project (Project No.: 2016GA049) “Research on Big Data-Driven High-risk Students Prediction”.

## REFERENCES

[1] Alejandro P. Education data mining. A survey and a data mining-based analysis of recent works[J]. Expert Systems with Applications, 2014,41(4):1432-1462.

[2] Romero C., Lopez M I, Luna J M, et al. Predicting students’ final performance from participation in online discussion forums[J]. Computer & Education, 2013,68:458-472.

[3] Natek S, Zwilling M. Student data mining solution-knowledge management system related to higher education institutions[J]. Expert Systems with Applications, 2014,(0).

[4] Caro E, Gonzalez C, Mira J M. Student academic performance stochastic simulator based on the Monte Carlo method[J]. Computer & Education, 2014,76(0):42-54.

[5] Lara J A, Lizcano D, Martinez M A, et al. A system for knowledge discovery in e-learning environment within the European Higher Education Area – Application to student data from Open University of Madrid, UDIMA[J]. Computer & Education, 2014,72:23-36.

[6] Hachey A C, Wladis C W, Conway K M. Do prior online course outcomes provide more information than GPA. Alone in predicting subsequent online course grades and retention? An observational study at an urban community college[J]. Computer & Education, 2014,72(0):59-67.

[7] Feldman J, Monteserin A, Amandi A. Detecting students’ perception style by using games[J]. Computer & Education, 2014,71(0):14-22.

[8] Arteaga Sanchez R, Cortijo V, Javed U. Students’ perceptions of facebook for academic purposes[J]. Computer & Education, 2014,70(0):138-149.

[9] Uddin S, Thompson K, Schwendiman B, et al. The impact of study load on the dynamics of longitudinal email communications among students[J]. Computer & Education, 2014,72(0):209-219.

[10] Chen X, Vorvoreanu M, Madhavan K. Mining Social Media Data for Understanding Students’ Learning Experiences[J]. Learning Technologies, IEEE Transactions on, 2014, PP(99):1-1.

[11] Hong J-C, Hwang M-Y, Liu M-C, et al. Using a “prediction-observation-explanation” inquiry model to enhance student interest and intention to continue science learning predicted by their Internet cognitive failure[J]. Computer & Education, 2014,72(0):110-120.

[12] Xiaogao Yu, Xiaopeng Yu. A new k-nearest neighbor searching algorithm based on angular similarity. ICMLC. Jul 12-15 ,2008:1779-1784.

[13] Kruger-Ross M J, Waters R D. Predicting online learning success Applying the situational theory of publics to the virtual classroom[J]. Computer & Education, 2013,61:176-184.

[14] Xiaogao Yu.The research on distributed adaptive text classification. WiCOM. Oct 12-14,2008:467-472.

[15] Yu Xiaogao. Research on international curriculum teaching [J], Journal of Higher Education, 2016,8:17-19.

[16] Yu Xiaogao. Research on the internationalization of management information system in the environment of big data [J].Software Guide, 2016,5:216-218.